# Enriching with Minority Instances a Corpus of Sentiment Analysis in Spanish

Juan-Luis García-Mendoza[1,*], Davide Buscaldi[1]

[1]Université Paris 13, Villetaneuse, France, 93430

Abstract

This paper addresses the challenge of enriching a Spanish sentiment analysis corpus with minority instances. Existing sentiment analysis resources for Spanish suffer from a significant class imbalance, limiting the representation of minority sentiments. We propose a methodology that leverages social media data and user-generated content to collect diverse Spanish text samples expressing a wide range of sentiments. We employ rigorous annotation and data augmentation techniques to ensure the quality and balance of the enriched corpus. Experimental results demonstrate that enriching the corpus with minority instances significantly improves sentiment analysis model performance, enhancing inclusivity and accuracy in Spanish sentiment analysis.

**Keywords**

Sentiment Analysis, NLP, BERT, data augmentation, Mexican tourism

## 1. Introduction

Sentiment analysis, also known as opinion mining, is a field of study that focuses on identifying and extracting subjective information from textual data. It plays a crucial role in various domains, including social media monitoring, market research, and customer feedback analysis [1]. Although sentiment analysis has gained significant attention in recent years [2, 3, 4, 5], most of the research and resources in this field are primarily focused on English language data [6]. This bias towards English limits the availability of sentiment analysis tools and resources for other languages, including Spanish [7].

Spanish is one of the most widely spoken languages globally, with a large and diverse population of native speakers [8]. It is essential to develop language-specific resources and models for sentiment analysis in Spanish to cater to the needs of Spanish-speaking individuals and businesses. Furthermore, sentiment analysis in Spanish poses unique challenges due to the rich morphology, complex syntax, and regional variations found within the language [9, 10].

In this paper, we propose an approach to enrich an existing corpus of sentiment analysis in Spanish by incorporating minority instances as part of the Rest-Mex Task 2023 [11]. The incorporation of minority instances is crucial to mitigate bias and enhance the generalizability of sentiment analysis models across diverse linguistic and cultural contexts. By incorporating

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

garciamendoza@lipn.univ-paris13.fr (J. García-Mendoza); davide.buscaldi@lipn.univ-paris13.fr (D. Buscaldi)

a diverse range of opinions, expressions, and sentiment nuances, we aim to develop a more robust and representative corpus for sentiment analysis in Spanish.

We present a methodology for collecting and annotating the minority instances in our corpus. Additionally, we discuss the potential applications and benefits of using this enriched corpus, including improving sentiment analysis models' performance, enabling more accurate opinion mining on social media platforms, and enhancing the understanding of sentiment trends across different demographic groups within the Spanish-speaking population.

## 2. Related Works

Sentiment analysis has been extensively studied in the English language, leading to the development of numerous resources, tools, and techniques. However, research on sentiment analysis in languages other than English is relatively limited, particularly for minority languages. In recent years, there has been an increasing interest in developing sentiment analysis resources for minority languages to bridge this gap [12, 13].

Several studies have focused on developing sentiment analysis resources and models for languages such as Arabic, Chinese, and Hindi [14]. These efforts have shown promising results in enhancing sentiment analysis performance in those languages. However, the specific linguistic characteristics and sentiment expressions in Spanish present distinct challenges that require dedicated research and resources.

Existing resources for sentiment analysis in Spanish are primarily based on small-scale corpora, often limited to specific domains or regions [2]. These resources lack diversity and fail to capture the full spectrum of sentiment expressions and nuances present in the Spanish language. Moreover, they do not adequately represent the opinions and sentiments of minority groups within the Spanish-speaking population.

To address these limitations, researchers have begun exploring methods to enrich sentiment analysis corpora with minority instances [15]. Incorporating diverse viewpoints, including those of underrepresented groups, helps to reduce bias and improve the generalizability of sentiment analysis models. However, these efforts are still in their nascent stages, and there is a need for more comprehensive and representative corpora for sentiment analysis in Spanish.

In this paper, we build upon existing research on sentiment analysis in Spanish by proposing a methodology to enrich an existing corpus with minority instances. Our approach takes into account the linguistic and cultural diversity within the Spanish-speaking

## 3. Dataset

The organizers of Rest-Mex 2023 have developed a comprehensive train collection consisting of 251,702 opinions sourced from TripAdvisor labeled for Polarity, Type and Country.

Regarding the Polarity classification, the dataset is divided into 5 classes. Class 1 represents the worst polarity, while class 5 denotes the best polarity. The distribution of these classes can be observed in Figure 1, which reveals a significant class imbalance.

To ascertain the Type of place, the dataset is classified into three distinct classes: Attractive, Hotel, and Restaurant. The distribution of this characteristic is depicted in Figure 2. Unlike the

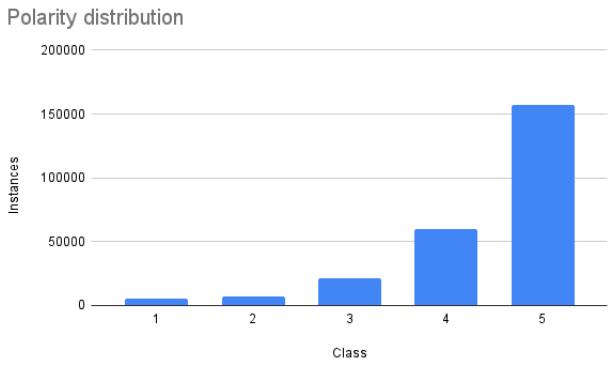


**Figure 1:** Polarity distribution

polarity classification, there is no explicit imbalance, but it is evident that the classes are not evenly balanced.

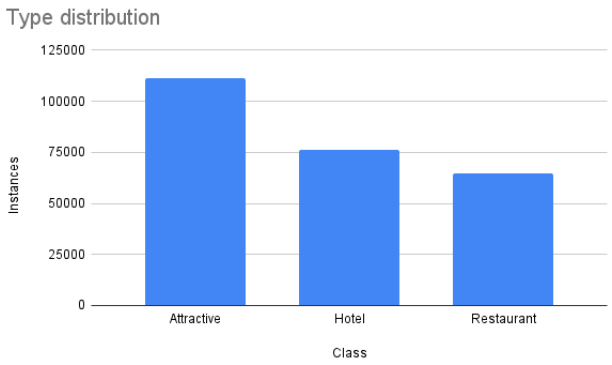


**Figure 2:** Type distribution

Lastly, the classification of the Country of origin for the places visited by tourists entails three classes within the collection: Mexico, Cuba, and Colombia. Figure 3 showcases the distribution of this attribute.

In this way, it can be seen that the trait with the least imbalance is that of Type, and that of polarity is significantly greater than the other two.

## 4. Methodology

This section presents the proposed methodology for enhancing the minority instances within the Rest-Mex 2023 database.

The methodology begins with a description of the text pre-processing techniques employed, followed by an outline of the added data to the collection. Subsequently, the classifier employed

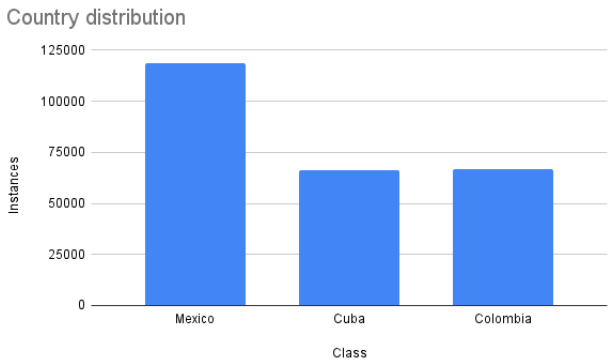


**Figure 3:** Country distribution

in the process is detailed, and finally, the evaluation metrics utilized are described.

### 4.1. Data pre-processing

In Section 3 we emphasize the convenience of doing some pre-processing to the data. Considering this, the steps for the pre-processing are:

- Lower case transformation
- Punctuation cleaning, unusual characters removal
- Repeated words
- Emojis replacement

### 4.2. Adding Negative Reviews to the Rest-Mex 2023 Collection

It is apparent that the minority classes align with the negative classes. Thus, the proposal is to incorporate instances representing these polarities.

Consequently, all the opinions classified as polarity 1 and 2 from the Rest-Mex 2022 collection [3] were extracted, resulting in a total of 1277 instances. Out of these, 547 opinions belong to polarity 1, while 730 opinions belong to polarity 2.

The objective is to determine whether the addition of a few instances can yield substantial outcomes for the minority classes.

### 4.3. Bert based Classification

To classify the data, it is proposed to use a transformer based on Bert but specialized for Spanish such as Beto.

Following we show the features used in the implementation used for the classifier:

- Model: Beto-cased
- Max length: 32

- Optimizer: Adam
- Learning rate: $5e-5$
- Steps: $1e-08$
- Epochs: 3

### 4.4. Metrics and Evaluation

The evaluation metrics introduced by the Rest-Mex organizers prioritize accurate classification of the negative polarity classes.

In order to gauge the efficacy of the polarity classifier, the organizers suggest utilizing Equation 1. This metric assigns a weight that is the additive inverse of the class instance percentage in the test collection, reflecting its relative importance.

$$Res_P(k) = \frac{\sum_{i=1}^{|C|}((1 - \frac{T_{C_i}}{T_C}) * F_i(k))}{\sum_{i=1}^{|C|} 1 - \frac{T_{C_i}}{T_C}} \tag{1}$$

For the evaluation of the Type and Country traits, the organizers propose the adoption of Equations 2 and 3. These metrics represent the macro F-measures for each respective trait.

$$Res_A(k) = \frac{F_A(k) + F_H(k) + F_R(k)}{3} \tag{2}$$

$$Res_C(k) = \frac{F_{Mex}(k) + F_{Cub}(k) + F_{Col}(k)}{3} \tag{3}$$

Lastly, to derive a single value per participant, the organizers suggest combining the results using Equation 4. Notably, it is worth mentioning that the polarity result carries more weight compared to the other two traits in this combined value.

$$Sentiment(k) = \frac{2 * Res_P(k) + Res_A(k) + Res_C(k)}{4} \tag{4}$$

This evaluation approach appears to be well-suited for the methodology proposed in this study, as it aims to achieve the optimal outcome while striking a certain balance within the corpus.

## 5. Results

For the train, we use a data split of the 70/30 partition. Table 1 shows the results obtained. The idea is to compare the result among the original data set versus the extended data set.

It is intriguing to note that the polarity results do not exhibit a substantial improvement, which could potentially be attributed to the evaluation being conducted on the same training corpus.

Table 2 presents the results achieved in the test partition of Rest-Mex 2023. It is noteworthy to observe the substantial increase in performance in the polarity classification compared to the

| Approach | F-Polarity | F-Type | F-Country |
|---|---|---|---|
| Original Data Set | 0.4927 | 0.9767 | - |
| Rest-Mex extended data set | 0.4985 | 0.9708 | 0.9089 |

**Table 1**
Train results

training partition. Additionally, there is an evident improvement in the results for the Type classification.

It is also important to mention that the results obtained exceed the 3 baselines proposed by the organizers.

What makes it even more intriguing is that this slight improvement is attained by adding only 1277 instances, which represent a mere 0.5% of the dataset. This suggests that further gains can be achieved by continuing to augment the dataset with additional instances.

| Approach | Sentiment Track Score | F-Polarity | F-Type | F-Country |
|---|---|---|---|---|
| Original Data Set | 0.4929 | 0.4464 | 0.9559 | 0.2137 |
| *Rest-Mex extended data set* | 0.6888 | 0.4818 | 0.9719 | 0.9081 |
| BseLine-Beto-No-Fine-Tuning | 0.3810 | 0.2476 | 0.8344 | 0.3474 |
| BaseLine-majority | 0.1407 | 0.1537 | 0.2047 | 0.2137 |
| BaseLine-minority | 0.1102 | 0.0092 | 0.2047 | 0.2137 |

**Table 2**
Test results

## 6. Conclusions

In this study, we presented a proposal to enhance the Rest-Mex 2023 database by incorporating data from another edition of the event.

The main objective was to acquire more instances of the negative classes, which, in the context of sentiment analysis in tourism, are recognized as the minority classes.

A total of 1277 instances from the 2022 collection were added, representing a mere 0.5% of the original dataset. Despite this small increase, a notable improvement in the F-measure was achieved, with an increase from 0.4464 to 0.4818.

These findings indicate that by introducing additional negative instances from other collections, it is possible to generate more effective classification models.

As future work, we propose exploring models that incorporate minority instances while also applying subsampling techniques to balance the majority classes more effectively.

## References


[1] E. Olmos-Martínez, M. Á. Álvarez-Carmona, R. Aranda, A. Díaz-Pacheco, What does the media tell us about a destination? the cancun case, seen from the usa, canada, and mexico, International Journal of Tourism Cities (2023).

[2] M. Á. Álvarez-Carmona, R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado, R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, A. Y. Rodríguez-González, Overview of rest-mex at iberlef 2021: recommendation system for text mexican tourism 67 (2021). doi:https://doi.org/10.26342/2021-67-14.

[3] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, D. Fajardo-Delgado, R. Guerrero-Rodríguez, L. Bustio-Martínez, Overview of rest-mex at iberlef 2022: Recommendation system, sentiment analysis and covid semaphore prediction for mexican tourist texts, Procesamiento del Lenguaje Natural 69 (2022).

[4] M. A. Alvarez-Carmona, R. Aranda, A. Rodriguez-Gonzalez, D. Fajardo-Delgado, M. G. A. Sanchez, H. Perez-Espinosa, J. Martinez-Miranda, R. Guerrero-Rodriguez, L. Bustio-Martinez, A. D. Pacheco, Natural language processing applied to tourism research: A systematic review and future research directions, Journal of King Saud University-Computer and Information Sciences (2022).

[5] A. Diaz-Pacheco, M. Á. Álvarez-Carmona, R. Guerrero-Rodríguez, L. A. C. Chávez, A. Y. Rodríguez-González, J. P. Ramírez-Silva, R. Aranda, Artificial intelligence methods to support the research of destination image in tourism. a systematic review, Journal of Experimental & Theoretical Artificial Intelligence (2022) 1–31.

[6] R. Guerrero-Rodriguez, M. Á. Álvarez-Carmona, R. Aranda, A. P. López-Monroy, Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico, Current issues in tourism 26 (2023) 289–304.

[7] M. A. Alvarez-Carmona, A. P. López-Monroy, M. Montes-y Gómez, L. Villasenor-Pineda, H. Jair-Escalante, Inaoe's participation at pan'15: Author profiling task, Working Notes Papers of the CLEF 103 (2015).

[8] M. Á. Álvarez-Carmona, E. Villatoro-Tello, L. Villaseñor-Pineda, M. Montes-y Gómez, Classifying the social media author profile through a multimodal representation, in: Intelligent Technologies: Concepts, Applications, and Future Directions, Springer, 2022, pp. 57–81.

[9] M. A. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso, L. Villaseñor-Pineda, Semantically-informed distance and similarity measures for paraphrase plagiarism identification, Journal of Intelligent & Fuzzy Systems 34 (2018) 2983–2990.

[10] M. Á. Álvarez-Carmona, R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, H. Carlos, Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news, Journal of Information Science (2022). doi:https://doi.org/10.1177/01655515221100952.

[11] M. Á. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Y. Rodríguez-González, L. Bustio-Martínez, V. Muñis-Sánchez, A. P. Pastor-López, F. Sánchez-Vega, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023).

[12] M. A. Alvarez-Carmona, R. Aranda, A. Diaz-Pacheco, J. de Jesús Ceballos-Mejıa, Generador automático de resumenes cientıficos en investigación turıstica, Research in Computing Science (2022).

[13] S. Arce-Cardenas, D. Fajardo-Delgado, M. Á. Álvarez-Carmona, J. P. Ramírez-Silva, A tourist recommendation system: a study case in mexico, in: Advances in Soft Computing: 20th Mexican International Conference on Artificial Intelligence, MICAI 2021, Mexico City, Mexico, October 25–30, 2021, Proceedings, Part II 20, Springer, 2021, pp. 184–195.

[14] H. Peng, E. Cambria, A. Hussain, A review of sentiment analysis research in chinese language, Cognitive Computation 9 (2017) 423–435.

[15] M. A. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodrıguez, A. Y. Rodrıguez-González, A. P. López-Monroy, A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one, Computación y Sistemas 26 (2022). doi:https://doi.org/10.13053/CyS-26-2-4055.